Promoting Effective Decision-Making: Training Educators to Collect and Use Social-Emotional Skill Assessment Data to Inform Tier II Interventions

Nathaniel von der Embse, Ph.D.

University of South Florida

Stephen Kilgus, Ph.D.

Katie Eklund, Ph.D.

Miranda Zahn, M.S.

University of Wisconsin-Madison

Casie Peet, M.A.

Sarahy Durango, M.A.

University of South Florida

School Psychology Review

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180515 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Correspondence regarding this article should be addressed to Nathaniel von der Embse at the University of South Florida, Department of Educational and Psychological Studies, Tampa, FL 33620; email: natev@usf.edu

Abstract

Teachers are often called upon to identify behavioral and emotional risk within their students by completing a variety of assessment tools. However, many teachers may lack the requisite skills to reliably identify students at-risk or use data derived from assessment tools to inform intervention. A series of trainings were developed to improve decision-making on the *Intervention Selection Profile—Skills,* with a focus on improving accuracy and use of data. Specifically, a two-study randomized controlled design was employed to evaluate the efficacy of a basic informational training and a training with a practice component with regards to a control condition on the collection and use of social-emotional assessment data on the ISP-Skills. Results suggest limited influence of training on the accuracy of data collection, yet significant influence on improving how data are used to inform intervention. Implications for practice and research, as well as limitations, are discussed.

Keywords: *assessment, social-emotional skills, teacher training, decision-making*

Promoting Effective Decision-Making: Training Educators to Collect and Use Social-Emotional Skill Assessment Data

Assessment and data-based decision-making remains a core component of effective educational service delivery. Nearly half of the typical roles and functions of school psychologists includes assessment related activities (Castillo, Curtis, & Gelley, 2013). However, the nature of assessment within schools is changing as practitioners are increasingly engaged in decision-making across tiers of service (Jimerson, Burns, & VanDerHeyden, 2007; Shapiro & Heick, 2004). School psychologists are now responsible for evaluating universal screening data at Tier I (Eklund, DeMarchena, Rossen, Izumi, Vaillancourt, & Kelly, 2019), determining selection and modification of group interventions at Tier II (McDaniel, Bruhn, & Mitchell, 2015), and engaging in more traditional individualized psychoeducational assessment for special educational classification at Tier III (Sullivan, Sadeh, & Houri, 2019). These expanding assessment roles necessitate (1) a change in conventional assessment training (Rossen & von der Embse, 2014), and (2) a modern approach to assessment research that goes beyond traditional psychometric evidence to identify the procedures through which assessment data can be effectively collected and analyzed across all tiers of service (von der Embse & Kilgus, 2018).

Robust psychometric evidence is one of several considerations when evaluating the defensibility of using an assessment tool for some applied purpose (e.g., universal screening, problem analysis, or progress monitoring). Other factors that influence defensibility include rater training, procedural guidance (e.g., frequency or timing of assessment), and guidelines for the interpretation of data to inform the decision. For example, research has begun to delineate essential procedures that improve the quality of decision making within academic curriculum-based measures (e.g., number of times per week, influence of goal line on intervention

determination; Christ, Zopluoglu, Monaghen, & Van Norman, 2013; Van Norman & Parker, 2018), and rater training on progress monitoring tools for behavioral intervention (e.g., collection schedules for direct behavior ratings [DBRs]; Chafouleas, Riley-Tillman, Jaffrey, Miller, & Harrison, 2015). New research has informed guidelines on how to interpret data to inform intervention (e.g., ordinate scaling on single-case graphs; Dart & Radley, 2017). These types of investigations that define evidence-based assessment *procedures* may help mitigate the likelihood of poor data input or inaccurate data evaluation. Given the overall lack of evidence to inform how assessments are used within applied contexts (e.g., schools), there have been repeated calls for novel research to inform assessment collection and use (VanDerHeyden & Burns, 2018; Volpe & Briesch, 2018). The aim of the present investigation is to describe one such effort designed to improve the quality of data to promote effective decision-making for intervention selection.

Assessment within Multi-Tiered Systems of Support

Multi-Tiered Systems of Support (MTSS) is an integrated educational framework that combines academic, behavioral, social-emotional, and expanding to mental health supports in increasing levels of intensity designed to meet the various needs of all students (Algozinne et al., 2012). The foundation of successful MTSS rests on the quality of data used to inform decisions across tiers of service. That is, for MTSS to be maximally effective and efficient, there is a need for quality data to (1) inform necessary modifications to Tier I (e.g., screening data to indicate disproportionality); (2) match instructional and/or contingency management interventions at Tier II; and (3) individualize intensive supports at Tier III (Arden, Gandhi, Edmonds, & Danielson, 2017). Traditionally, MTSS frameworks have recommended a continuum of intervention supports in increasing intensity *across* rather than *within* Tiers I, II, and III (Kilgus, Collier-

Meek, Johnson, & Jaffery, 2014). For example, many schools have a universal curricula or set of practices implemented at Tier I (e.g., Positive Behavior Interventions and Supports [PBIS]) and a standard protocol approach to Tier II interventions (e.g., Check In, Check-Out, [CICO]; Bundock, Hawken, Kladis, & Breen, 2019); these practices typically relegate the individualization or modification of interventions to Tier III. However, calls have been made for problem solving approaches to be incorporated at Tier II decision-making thus improving the likelihood of intervention match and subsequent effectiveness (Reinke, Stormont, Clare, Latimore, & Herman, 2013).

Researchers have begun to provide guidance by which to support the individualization of interventions across both Tiers II and III (Kilgus & von der Embse, 2019). In accordance with a data-based individualization model (Kuchle, Edmonds, Danielson, Peterson, & Riley-Tillman, 2015), researchers have proposed a hybrid intervention approach, wherein (1) a student enters into a standard protocol Tier II intervention (e.g., CICO) after being identified as at-risk via universal screening; (2) progress monitoring is employed to determine intervention response; and (3) if the student is non-responsive, problem analysis is conducted to determine a student's specific needs. The problem analysis data then support the selection of intervention strategies aligned with the student's concerns. Previous research has supported the hybrid intervention approach, suggesting that while standard protocol interventions will be effective for many students (Hawken & Horner, 2003), modifications to these interventions can enhance their effectiveness for students who are otherwise non-responsive. Such modifications include function-based adaptations of Check In/Check Out (Kilgus, Fallon, & Feinberg, 2016), as well as targeted social-emotional skill instruction targeting a student's specific skill deficits (Barreras, 2009). Unfortunately, research is lacking regarding efficient and feasible problem analysis tools

that can support this intervention individualization at Tier II (Bruhn, Wehby, & Hasselbring, 2019). The dearth of problem analysis tools suitable for use at Tier II is especially pronounced within the social-emotional domain.

Social-Emotional Skills Assessment within MTSS

Social-emotional skill assessments afford information that can guide behavioral skill training interventions. The assessments suggest which skills a student has yet to acquire and should therefore be targeted for behavioral skills training (Elliott, Gresham, Frank, & Beddow, 2008). There are a number of social-emotional skill assessment tools with substantial evidence for use at the individual student level (e.g., traditional rating scales, semi-structured interviews, structured observations; Whitcomb, 2013). For example, the *Social Skills Improvement System* (SSIS; Gresham & Elliott, 2008) is a suite of tools that features teacher, student, and parent rating forms with 46 items (teacher form) that identify student functioning in seven social skills domains. The SSIS has strong psychometric support and demonstrated treatment utility (Elliott et al., 2008). Nevertheless, use of the SSIS (and other measures of its length) can prove challenging at Tier II given the time and effort associated with its use. The developers suggest the SSIS takes 10-15 minutes to complete per student (Gresham & Elliott, 2008). Given that nearly 20% of students need Tier II supports (Schanding & Nowell, 2013), it would take almost an hour in a typical classroom of 25 students to complete the number of Tier II assessments necessary to inform intervention grouping, and even more time to use for progress monitoring purposes. A more efficient, and ultimately more feasible, assessment process would necessitate briefer skill assessments, allowing for more widespread use at the Tier II scale (Kilgus & von der Embse, 2019).

Researchers have begun to answer the call for brief and efficient skill assessments. For instance, Kilgus, Eklund, and von der Embse (2018) recently developed the *Intervention Selection Profile – Skills* (ISP-Skills), a brief 14-item teacher rating scale designed to assess a range of social-emotional skills (e.g., self-awareness and relationship skills) and academic enabling skills (e.g., motivation and academic engagement). The ISP-Skills is built on diagnostic classification modeling (DCM), a confirmatory multidimensional latent-variable modeling approach (Rupp & Templin, 2008). DCM emphasizes *within-item* multidimensionality, such that each item is modeled as providing information about one or more discrete latent attributes. This stands in contrast to the more familiar concept of *between-item* multidimensionality, which is common in factor analysis and related methods, which aims to identify homogeneous clusters of items or indicators. Thus, DCM posits that an item set may be unidimensional in the traditional sense (i.e., all items may measure a single underlying construct), yet multidimensional due to the combinations of attributes that characterize each item. Accordingly, through DCM-based scoring, an item can provide information about multiple attributes, permitting an abbreviated measure to still provide information about a wide range of attributes.

Though similar to item response theory (IRT), DCM differentiates itself in terms of the manner in which it conceptualizes latent attributes. Within IRT, latent variables represent continuously scaled estimates of student ability, which may be interpreted in a manner consistent with *z* scores ($M = 0$, $SD = 1$). Within DCM, latent variables represent estimates of the probability of a certain attribute being present, scaled from 0-1. Such probability-based information is of value within the context of a skill assessment, where the goal is to determine which skills a students have yet to acquire and should therefore be targeted for instruction.

In accordance with the DCM scoring approach, the ISP-Skills yields a series of scores indicative of the probability a student has mastered eight skills. A low probability score would suggest the presence of a skill deficit that should be targeted for behavioral skills training. A recent study supported the performance of ISP-Skills items and subscales, with findings revealing high levels of internal consistency reliability (α = .93-.94), concurrent convergent validity relative to a range of criterion measures (e.g., the SSIS; r = .70-.86), and diagnostic accuracy in predicting the presence of below average skills (area under the curve = .84-.92; Kilgus et al., 2020).

Though promising, this initial study only supported the psychometric defensibility of ISP-Skills scores. Moving forward, researchers should examine necessary procedures, including rater training, that are needed to ensure the quality of data going into the tool as well as how data are interpreted to inform intervention selection. For example, multiple studies have indicated rater training, paired with opportunities to practice with performance feedback, have improved the accuracy of student behavior ratings used for the purpose of progress monitoring or functional behavior assessment (Chafouleas, Riley-Tillman, Jaffrey, Miller, & Harrison, 2015; Kilgus, Kazmerski, Taylor, & von der Embse, 2017). Chafouleas and colleagues (2015) utilized a multicomponent online training module designed to improve a rater's construct knowledge inclusive of practice opportunities. Although results were promising, limitations included use of undergraduate students and need to determine initial construct knowledge. Building on this work, Kilgus and colleagues (2017) used multiple training conditions to ensure consistency in initial construct knowledge, while varying exposure to amount and type of rater training. However, undergraduates were used as a convenience sample (Kilgus et al., 2017). Further research with classroom teachers has supported the initial efficacy of rater trainings that improve predictability

of screening with a number of proximal and distal student outcomes (von der Embse, Kilgus, Eklund, Ake, & Levi-Neilsen, 2018). However, there is limited knowledge of key processes that may influence skill assessment rating accuracy and no empirical investigations to determine the extent to which training may influence the decisions based upon assessment data. Overall, there is a clear need for additional research to inform critical components of effective rater training protocols.

Purpose of the Investigation

To fully realize the promise of Tier II interventions within MTSS frameworks, a novel approach to Tier II assessment is necessary. There is a critical need to delineate key procedures that may influence the quality of data entering into an assessment process. Given the assumptions made within many MTSS assessment frameworks on rater construct knowledge and how data are used, there is a potential of inaccurate and inefficient decisions (Evans et al., 2005). In addition, there is a need for modern psychometric approaches to consider the social validity, or impact of the decisions made from assessment data, when evaluating a tool's evidence for use. As such, the current investigation seeks to evaluate evidence-based training methodology within Tier II social skills assessment that may ultimately support use within school contexts.

Two research questions were of interest. First, what level of educator training is necessary to support the *collection* of accurate ratings of student social-emotional functioning on the *ISP - Skills*? Based upon previous research (e.g., Kilgus et al., 2017), it was hypothesized that some level of basic rater training along with opportunities to practice rating students would be necessary to support rater accuracy. Second, what level of educator training is necessary to support the accurate *use* of *ISP – Skills* ratings regarding student social-emotional functioning? In accordance with prior findings (Loman & Horner, 2014), it was expected training and practice

would be necessary to support educators in using social-emotional skill data to make accurate intervention-related decisions. Two studies were conducted to examine these two research questions. Both studies represented randomized controlled trials (RCTs), wherein educator participants were randomly assigned to one of three conditions: Control, Basic Training, and Basic Training + Practice. Conditions were compared with regard to their capacity to support educators in collecting accurate ratings of hypothetical students described in vignettes (Study 1) and deriving accurate decisions regarding appropriate intervention strategies in consideration of hypothetical data reports (Study 2).

STUDY 1

Method

Participants

Study participants included educators sampled from two elementary schools (grades K-5) within a single suburban school district in the upper Midwest. The majority of participants who provided role information were general education teachers, while a subset were administrators (with previous teaching experience), special education teachers, or student support personnel (e.g., school psychologists and counselors). A detailed breakdown of the participant demographics is presented in Table 1. Overall, 96 educators participated in this study, with 47 being from School 1 and 49 from School 2. Of these 96 educators, 37 were randomly assigned to the Control condition, 31 to the Basic Training condition, and 28 to the Basic Training + Practice. The majority of educators were female (90.6%; 9.4% male) and White (99.0%). (One educator indicated they did not wish to provide their race/ethnicity). Regarding experience in education, 21.9% of participants had 0-5 years, 24.0% has 5-10 years, and 54.2% has 10+ years.

An *a priori* power analysis was conducted using the G*Power software to inform Study 1 recruitment. The power analysis was specific to a repeated measures MANOVA examining the interaction between within-group (i.e., Time) and between-group (i.e., Condition) factors. The power analysis assumed two repeated measurements, three group conditions, a critical alpha level of .05, and an effect size *f*(V) equal to 0.29. This anticipated effect represented the mean of effect sizes corresponding to the interaction effects from repeated measures MANOVA tests reported in Kilgus et al. (2017). Results suggested a sample of 114 participants would be necessary to achieve power of .80. We acknowledge that while the current sample approximated this total number, it still fell slightly short of the recommended sample size.

Measures

Intervention Selection Profile. The ISP-Skills (Kilgus et al., 2018) is a brief teacher rating scale (14 items), designed for use in problem analysis with students identified for intervention. Five of the skills within the ISP-Skills are subsumed under the Social-Emotional domain. These five skills are aligned with the Collaborative for Academic, Social, and Emotional Learning (2005) Core Five Competencies, which include Self-Awareness, Social Awareness, Relationship Skills, Self-Management, and Responsible Decision Making. The remaining three skills within the ISP-Skills correspond to the Academic Enablers domain (DiPerna, 2006), and include Study Skills, Academic Engagement, and Motivation. Each ISP-Skills item is completed using a behaviorally anchored rating scale (BARS), which includes five anchors corresponding to particular skill levels defined by level of *skill acquisition* (i.e., the degree to which the skill has been learned) and *skill utilization* (i.e., the degree to which the skill is used once learned). When taken together, the anchors represent the categories of skill development commonly conceptualized in research and practice, including *acquisition deficit, fluency deficit,*

performance deficit, typical, and *strength* (Gresham, Elliott, & Kettler, 2010). See Table 2 for operational definitions of these terms.

Vignettes. To address the central purpose of Study 1, it was necessary to create a measurement scenario within which the accuracy of ISP-Skills ratings could be judged. Evaluations of rating accuracy necessitate some standard or "true score" against which scores can be compared (Guion, 1965). Observed scores closer to a true score are considered more accurate, while those farther away from a true score are considered less accurate. Under typical circumstances, educators would use the ISP-Skills to rate the behavior of a student with whom they are quite familiar. Unfortunately, such usage is challenging with rater training studies, as it is difficult (if not impossible) to derive true scores for all students that might be considered. Though it would be helpful if all participants could rate the same subset of students for which true scores could be reasonably derived, it is unlikely all participants will be familiar with these students. Accordingly, the decision was made to have all participants rate the same set of hypothetical students described through a series of nine vignettes, including three pretest vignettes, three practice vignettes, and three posttest vignettes.

The research team developed the vignettes, which described hypothetical students within a variety of social and academic settings. Each vignette depicted a target student engaging in certain social-emotional or academic enabling skills with specific degrees of fluency and frequency. Vignettes were specifically designed to introduce variation in the skills described, as well as the degree of fluency and frequency with which target students exhibited skills. This approach was employed to ensure the study evaluated rater accuracy in relation to a number of rating scenarios. It was intended that study participants would read these vignettes and then complete two ISP-Skills items aligned with the skills described in the vignette. Our focus on only

two items per vignette is founded in two factors. First, our central intent in this study was to evaluate the accuracy with which raters used the BARS to rate ISP-Skills items, and not necessarily their capacity to differentiate amongst various skills when rating a broader set of items. Second, given our interest in developing brief vignettes, passages could only include information related to a narrow range of skills without becoming confusing or cumbersome.

A post-doctoral researcher with substantial experience and expertise in social-emotional and behavioral assessment, as well as creative writing, developed initial drafts of the nine vignettes. The first, second, and third authors then reviewed these drafts and provided edits and feedback. The vignettes were then randomly assigned to be used as part of pretest, practice, or posttest within study sessions (three vignettes per phase). On average, vignettes included 306.78 words (*SD* = 75.79, Range = 168-389) and 23.33 sentences (*SD* = 8.20, Range = 9-33), and were written at a grade level of 5.58 (*SD* = 1.23, Range = 3.97-8.09), per the Flesch-Kincaid index. See Appendix A for an example of one vignette (all vignettes are available from the first author upon request).

Prior to the study, a panel of individuals with advanced expertise in social-emotional and behavioral assessment were convened to generate true scores for each item corresponding to each vignette. The five panel participants included one school psychology professor and four advanced school psychology doctoral candidates, all of whom had completed all program coursework (including graduate coursework and practicum experience in social-emotional and behavioral assessment), and were preparing for their internship the following year. The panel session began with the session facilitator (i.e., the second author) providing an overview of the ISP-Skills, including information related to its purpose, items, factor structure, and the behaviorally anchored rating scale. The panelists were then given an overview of this study and

the role the vignettes would play within the investigation. Next, panelists reviewed each vignette and generated true scores as part of an expert consensus process, which has been used within other social-emotional and behavioral assessment studies (e.g., Kilgus et al., 2015).

First, all experts independently reviewed the first vignette. Once they had finished reading the vignette, the experts completed their ratings for the two ISP-Skills items corresponding to that vignette. More specifically, the experts selected the BARS anchor they felt most accurately described the skills the target student exhibited within the vignette. Each expert's ratings remained private until all experts were finished. Second, the panelists shared their ratings for each of the two ISP-Skills items. No further discussion of an item was necessary if experts were in agreement. If a level of disagreement was noted, the experts were given the opportunity to discuss their particular ratings. During discussions, experts typically provided a rationale for their ratings, while also justifying why they did not select other possible ratings. Experts also occasionally expressed uncertainty in their ratings, indicating they did and could still consider other options. The session facilitator made no comment throughout the majority of the discussion but would occasionally summarize or seek clarification from experts if necessary. Third, once the discussion appeared to reach a natural conclusion, the session facilitator asked if any panelist would like to change their ratings. Revisions were not required and were only carried out if at least one expert expressed interest in rating revision. Once the revision process began it proceeded in accordance with the first step described above. Experts could choose to either make the same ratings as before or choose a new rating. The experts then moved on to examine the next vignette once the re-rating process was complete. Across each vignette, perfect agreement was achieved among experts for all but two ISP-Skills items. For each of these items, a single expert continued to select a rating that differed from the other four experts' ratings. For

items with perfect agreement, true scores corresponded to the unanimously selected item rating. For the two items evidencing disagreement, the true score corresponded to the rating selected by the majority of experts on the panel.

Procedures

All study procedures were approved by a university Institutional Review Board (IRB) prior to the development of measures or training implementation. All procedures described below were conducted during a single session at each school during a meeting held prior to the start of a school day. Prior to the meeting, a school administrator had (1) informed all educators at the school that the research team would be present during the meeting, (2) provided a brief overview of the purpose of the study, and (3) given educators details related to nature and duration of study activities. Each meeting then began with a brief introduction of the research team and a review of the study and its objectives. Educators were then given the opportunity to consent to participating in the study. If educators consented, they signed and immediately returned a consent form to the researchers. If they did not consent, they were given options for alternative activities to engage in while the research activities were conducted.

Next, consenting educators were randomly assigned to conditions in accordance with this randomized controlled trial design. Researchers accomplished this by providing each participant with a slip of paper from a shuffled pile. Each slip of paper included a number (1-3) corresponding to a specific experimental condition. Participants were instructed to move to a different room within the school corresponding to their particular condition number. The remaining study procedures were then carried out within these rooms. All study condition sessions were run by either a school psychology doctoral student, post-doctoral researcher, or professor. All sessions were highly structured, incorporating a PowerPoint presentation that

depicted key points and concepts, as well as detailed scripts to which session leaders adhered during their presentation. Integrity checks were completed at 100% completion across conditions on both studies with a self-report checklist. (All PowerPoint presentations and scripts are available from the first author upon request.)

Three experimental conditions were compared within this study, including Control, Basic Training, and Basic Training + Practice. All three training conditions were completed in less than 60 minutes. Participants in all three conditions completed the same pre and posttests, each of which included three standard vignettes. The remaining procedures then varied by condition.

Within the Control condition, sessions began with participants completing the pretest, wherein they reviewed three vignettes and rated the two items corresponding to each. Next, the participants completed the posttest, which involved the review and rating of three additional vignettes. No training was provided between completion of these two test sets, suggesting any change to be noted between them in rating accuracy would be due to practice effects alone. Following posttest, Control condition participants were provided a basic training on the ISP-Skills. The effect of this training was not evaluated for this condition. Rather, the training was provided to ensure all participants received some information regarding the ISP-Skills. The basic training began with a description of how students use a variety of social-emotional and academic skills to complete daily living tasks. It was further noted that for any given skill, students can vary in terms of their acquisition and utilization of that skill. *Acquisition* was defined as the degree to which a person has learned a new skill, as indicated by the fluency of its display relative to normative expectations. *Utilization* was defined as the degree to which a person actually uses the skill after having learned it. Participants were informed that when considering these characteristics in tandem, one could differentiate student skill performance in terms of five

levels: acquisition deficit, fluency deficit, performance deficit, typical, and strength (operational definitions were provided for each of these levels). Session facilitators then reviewed how the ISP-Skills BARS anchors mapped on to each of the levels, and thus how item ratings are intended to provide information related to a student's various skill levels. Next, participants were provided an overview of the attributes the ISP-Skills is intended to assess, including the five social-emotional and three academic enabler attributes. The training then ended with a review of each ISP-Skills item. Participants were provided a detailed description of each item, along with examples of item-aligned behaviors that students might display within the school setting.

The Basic Training condition began with the pretest, followed by the same Basic Training described above. Sessions then ended with the posttest. The ordering of events within this group was such that any change in rater accuracy between pre and posttest was due to practice effects, as well as the influence of Basic Training.

Finally, the Basic Training + Practice condition began with the pretest followed by the Basic Training. Participants then practiced completing ISP-Skills items across three separate vignettes, which were distinct from those completed during either pre or posttest. For each practice opportunity, the participants quietly read the vignette and then completed the two ISP-Skills items aligned with the vignette. The session facilitator then provided participants with each item's true score for that vignette, which had been derived through the expert consensus process. A brief justification for the true score was provided and participants were given the opportunity to ask clarifying questions. Basic Training + Practice sessions then concluded with the posttest. To note, all training materials and vignettes are available from the first author upon request.

Data Analysis & Results

Prior to analysis, all variables were checked for data entry errors and missing data. No missing data were identified. For Study 1, it was necessary to create difference scores indicative of the accuracy with which participants rated pre and posttest vignettes. In accordance with prior research (e.g., Chafouleas et al., 2012; Kilgus et al., 2017), difference scores represented corrected absolute comparison scores (AB), defined as the absolute value of the difference between a participant's rating and the true score ($AB = |x_{true} - x_i|$). AB scores closer to zero were indicative of greater accuracy, and difference scores had a possible range of 0 to 10 because raters had the potential to be in exact agreement with the expert codes or any distance from true scores within the 10-point range. The mean of AB scores was then calculated within the pre and posttest phases, supporting the derivation of two scores for each participant that were then considered within the analyses described below. Descriptive statistics are presented in Table 3.

Parametric assumptions were examined for AB scores via Shapiro-Wilk tests, skewness and kurtosis statistics, Q-Q plots, boxplots, and histograms. Pre and posttest scores for the control had outliers that caused violations to normality. Outliers (n = 3) were removed for analysis, and a comparison between analyses revealed that reported trends were not influenced by the removal of outliers. Therefore, outliers were retained for all reported analyses. Variables violating the normality assumption are indicated in Table 3. A summary of AB score descriptive statistics for Study 1 is presented in Table 3. No significant differences in pretest AB scores emerged in pairwise comparisons with Bonferroni adjustment (ps > .585), providing evidence for baseline equivalence.

A repeated-measures MANOVA was used to examine differences between the three group conditions. Of interest within this analysis were the (a) main effect for time (within-subjects factor), with levels including pretest and posttest, (b) main effect for group (between-

subjects factor), with levels representing the three experimental conditions, and (c) interaction between group and time, which indicated the extent to which pretest-posttest score changes differed between groups. Pairwise comparisons are reported in Table 4. The interaction effect was non-statistically significant, indicating there was no difference in pretest-posttest score change between conditions, Pillai's Trace $F[2,89] = .415$, $p=.662$, partial $\eta^2 = .009$. However, the main effect for time was statistically significant, suggesting all conditions demonstrated positive growth in accuracy, as demonstrated by decreases in mean difference scores from pre to posttest, Pillai's Trace $F[1,89] = 41.653$, $p<.001$, partial $\eta^2 = .319$). Pairwise comparisons provided converging evidence for this finding, as there were no significant differences between the accuracy of ISP-Skills ratings at post-test between training groups.

STUDY 2

Method

Participants

In the second study, a sample from a Southeastern state included 198 educators who were part of their school-based problem-solving team. See Table 1 for a detailed breakdown of the participant demographics. The majority of participants were female (91.4%), White (80.3%), and non-Hispanic or Latino (82.8%). General education teachers (58.1%) and Exceptional Student Education (ESE) teachers or instructional coaches (18.2%) made up the majority of participants for a total of 76.3% of the sample. In addition to teachers, support staff (i.e. instructional assistants), student services personnel (e.g., psychologists, social workers, behavior specialists, and counselors) and administrators (e.g., principals and assistant principals) made up the remaining part of the sample. Participant years of experience varied with a little over a third within their first 5 years in the role (37.9%). All but 2.5% of the sample had at least a Bachelor's

degree. The power analysis conducted to inform Study 1 was also used to inform recruitment for Study 2. Accordingly, the sample for Study 2 met the recommended sample size.

Measures

Vignettes. All participants completed a pre-test at the beginning of each training session. The pretest included a brief demographic form to collect information on the participant's current position, number of years in that role, highest degree attained, gender, race, and ethnicity. The pre- and post-tests were made up of a series of vignettes on each assessment. Each vignette included a brief description of a hypothetical student and problem behavior, with a graph depicting the results from the ISP-Skills. All vignettes were fictitious but were intended to represent typical cases seen in schools. Descriptions of the students were intended to be brief and provide a school context for the presenting behaviors. All descriptions included gender-neutral terms and did not include any demographic information about the student to avoid potential biases. The ISP-Skills graph displayed the results of the ISP-Skills rating scale on each of the eight skills in terms of probability of mastery. This graph used a probability scale from 0-1 with 0 being a likely *deficit*, 1 being likely *mastered,* and 0.5 being 'undetermined'. Vignettes illustrated students with no skill deficits, one deficit, or two deficits.

Each of the vignettes had "complete the statement" type questions with multiple choice options (see Figure 1). Participants were prompted to select all that apply or to leave it blank if the question did not apply. The first question was, *"Student has a deficit in…"* followed by each of the skill areas. For example, if the participant interpreted that the student had a skill deficit in motivation and study skills, they selected only 'motivation' and 'study skills'. Participants had to determine how many (if any) of the students' score qualified as a deficit warranting intervention.

The final question was intended to ensure participants were able to link interpretation to intervention selection by prompting, *"The appropriate next step would be to…"*.

Participants selected from three multiple choice options for appropriate interventions. This last step was to ensure teachers and staff were making the connection between data interpretation and next steps as intervention selection is the intended purpose of the tool. The ISP-Skills is used to inform intervention selection, and more specifically, guide instructional targets. A pre and post-assessment were created to evaluate participant growth and mastery of data interpretation. The research group developed a series of vignettes that would evaluate the participants' ability to interpret the data. Each vignette had a 'profile' for the correct response pattern (e.g., two clear skill deficits and inconclusive functional data). For every vignette on the pre-test there was a vignette on the posttest with a matching profile to ensure a continuity of difficulty across the two assessments. After the assessments were developed, they were reviewed by a team of behavioral assessment experts to ensure agreement of the correct response profile. All vignettes used in the assessments received 100% agreement across the four developers and three outside experts. The vignettes were also reviewed by non-experts (i.e., teachers) for feedback on difficulty, layout, and general ability to complete the assessment.

Procedures

A 'Basic Training' protocol was created to teach how to use the problem-solving model and data-based problem analysis to target skill deficits when developing skill interventions. The training described how the ISP-Skills can be utilized within this process and the data elicited from these tools. In addition, the training provided explicit instructions for interpreting various data profiles and how to determine if there is a skill deficit. Lastly, the training included instruction on how to target the most likely skill deficit or function of behavior with either a

skills training or functionally adapted contingency intervention. A second training protocol, 'Basic Training + Practice' added a 'practice' section with example data profiles and performance feedback on participants' interpretation. Graduate students were trained in each of the training protocols and served as trainers for teachers and staff. Each of the training conditions were completed in approximately one hour.

The research team partnered with a large school district in the Southeastern United States. Following the approval of both the University IRB as well as the IRB in the partnering school district, the team began school recruitment. Trainings were scheduled by principals during a teacher in-service day, during a staff meeting, or other non-student day. Upon arrival at the school, participants were randomly assigned to one of three training conditions, which all took place at the same time in different rooms. There were approximately the same number of participants in each condition. In the Control Condition, participants took both the pre-test and the post-test at the beginning of the training. This group received the Basic Training following the post-test. This was in order to ensure the value of the training to the schools, rather than only training two-thirds of participants. Both the treatment groups (Basic Training and Basic Training + Practice) received training following the completion of the pre-test. After the training was delivered, participants completed the post-test and received a gift card for their participation.

Data Analysis & Results

Prior to analysis, all variables were checked for data entry errors and missing data as well as testing assumptions with a similar analytic plan as Study 1. Scores used for analysis in Study 2 were the proportion of correct decisions made rather than the AB score. To calculate this score, all decisions were coded as either in or out of agreement (1 or 0, respectively) with the true

score. Then, the proportion of correct decisions was calculated as the mean of correct and incorrect scores.

All variables violated assumptions of normality, which may be due to the restricted range of responses and a ceiling effect. As a precaution, the Pillai's Trace test statistic is reported for subsequent analyses given its robustness to the violation of parametric assumptions (Tabachnick & Fidell, 2007). A summary of comparison score descriptive statistics for Study 2 is presented in Table 5. With regard to the proportion of skills identified correctly by participants, no statistically significant differences emerged in pairwise comparisons with Bonferonni adjustment for pre-test scores for ISP-Skills overall, Motivation, Academic Engagement, Study Skills, Relationship Skills, Self-Awareness (*p*s = 1.000), Responsible Decision Making (*p*s > .689), Self-Management (*p*s > .703), Social Awareness (*p*s > .647).

As with Study 1, a repeated-measures MANOVA examined main effects for condition and time, as well as the interaction between condition and time. The interaction effect was found to be statistically significant, suggesting the groups differed with regard to their growth in accuracy from pretest to posttest, Pillai's Trace $F[2,195] = 15.365$, $p < .001$, partial $\eta^2 = .136$. With regard to ratings of specific skills, a main effect of training condition emerged for Motivation (Pillai's Trace $F[2,195] = 5.405$, $p = .005$, partial $\eta^2 = .053$), Academic Engagement (Pillai's Trace $F[2,195] = 3.675$, $p = .027$, partial $\eta^2 = .036$), Responsible Decision Making (Pillai's Trace $F[2,195] = 6.359$, $p = .002$, partial $\eta^2 = .061$), Relationship Skills (Pillai's Trace $F[2,195] = 22.348$, $p < .001$, partial $\eta^2 = .186$), Self-Management (Pillai's Trace $F[2,195] = 3.064$, $p = .049$, partial $\eta^2 = .030$), and Social Awareness (Pillai's Trace $F[2,195] = 18.535$, $p < .001$, partial $\eta^2 = .160$). There was no main effect of training condition on the proportion of

accurate ratings for Study Skills (Pillai's Trace F[2,195] = 2.509, p = .084 , partial η^2 = .025) or Self-Awareness (Pillai's Trace F[2,195] = 2.902, p = .057, partial η^2 = .029).

With regard to ratings of specific skills, a main effect of time emerged for Skills Overall (Pillai's Trace F[1,195] = 106.94, p <.001, partial η^2 = .35), Motivation (Pillai's Trace F[1,195] = 55.861, p <.001, partial η^2 = .223), Academic Engagement (Pillai's Trace F[1,195] = 119.311, p <.001, partial η^2 = .380), Study Skills (Pillai's Trace F[1,195] = 41.726, p <.001 , partial η^2 = .176), Responsible Decision Making (Pillai's Trace F[1,195] = 4.187, p = .042, partial η^2 = .021), Relationship Skills (Pillai's Trace F[1,195] = 113.667, p < .001, partial η^2 = .368), Self-Management (Pillai's Trace F[1,195] = 64.278, p <.001, partial η^2 = .248). There was no main effect of time on the proportion of accurate ratings for Social Awareness (Pillai's Trace F[1,195] = 2.744, p = .099 , partial η^2 = .014). or Self-Awareness (Pillai's Trace F[1,195] = .370, p = .544, partial η^2 = .002.

See Table 5 for a summary of pairwise comparisons. Pairwise comparisons indicated that the Basic Training and Basic Training + Practice conditions result in a higher proportion of accurate ratings at post-test for all variables compared to the control condition (ps ≤ .002). However, for nearly all skills, there was no discernible difference between Basic Training and Basic Training + Practice. For example, significant pairwise comparisons between the training conditions only emerged for Motivation (p = .024) and Relationship Skills (p = .004) with the Basic Training + Practice condition performing slightly better.

Discussion

The nature of assessment practices and decision-making more broadly within school settings is evolving (Castillo et al., 2013). As schools are adopting MTSS frameworks across the country, there is a critical need for responsible and defensible assessment to inform efficient

decision-making across Tiers of service. Much of the data collected across Tiers is dependent on classroom teachers; however, knowledge of constructs or how to use data should not be assumed particularly in the social-emotional and behavioral domains (Evans et al., 2005). Thus, a primary goal of the present investigation is to evaluate novel training methods to improve the quality of data entering into and decisions derived from a Tier II assessment tool. Two studies across two different sites were conducted to determine the relative influence of basic training and basic training + performance feedback on the accuracy of ratings of student behaviors (Study 1) as well as accurate decisions derived from data to drive intervention strategies (Study 2).

Results from the first study (i.e., data collection) suggested that participants had a similar baseline level of construct knowledge across the social-emotional skills domains measured on the ISP-Skills. When comparing the treatment groups to the control group at post-test, there were no significant differences noted. That is, irrespective of training modality, participants generally indicated a similar level of accuracy in identifying social-emotional skill deficits on the *ISP - Skills*. These preliminary data based upon rating vignettes indicate teachers may be relatively accurate (in comparison to other rating domains) in identifying areas of needs when utilizing the *ISP - Skills*. However, given the high accuracy scores on the pre-test, there may have been a ceiling effect that resulted in limited room for growth. In addition, caution is urged in generalizing these results given the analog rating scenario (vignettes) and need for research within naturalistic settings. Alternatively, and compared to rating internalizing types of behavioral concerns (Cunningham & Suldo, 2014), teachers may be more accurate in identifying social-emotional skills that could be more easily observed in a typical classroom environment. These results also provide initial evidence that the *ISP – Skills* may be completed by teachers without extensive rater training or professional development. However, additional research will

be necessary to determine initial rater construct knowledge with schools that have varying levels of social-emotional assessment use.

In comparison to the first study, the second investigation examined how teachers interpreted and used data derived from the *ISP-Skills*. Participants were presented with a series of graphs and asked to interpret skill deficits as well as identifying the necessary next steps. Results suggested that both the basic training in data interpretation and use, and the basic training with performance feedback resulted in increased accuracy in all of the eight domains, yet there was no significant difference between treatment conditions. These findings indicate the potential benefit of training in how to interpret data, yet performance feedback did not yield significant gains. Members of the problem-solving team play a critical role in evaluating data to inform intervention; results from Study 2 support the potential value of training these team members on how to interpret graphs as well as necessary next steps (i.e., intervention selection, or more data required).

Implications for Practice

Current results offer several considerations for practice. First, Study 1 results suggest that educators may have knowledge on basic behavioral constructs that would be necessary to complete the *ISP - Skills*. Given the preliminary results of Study 1, there does not appear to be a need for additional training for educators to derive accurate conclusions regarding student skills specifically on the *ISP - Skills*. Although rater variance may be present when there is a lack of construct knowledge (Evans et al., 2005), this implication aligns with general guidance that is provided to educators using similar skill-based measures (e.g., DiPerna, 2006) suggesting that as long as the rater has sufficient knowledge of the child, they may be able to complete these types of measures in the absence of any specialized training.

In contrast, Study 2 results suggest there is a benefit of training educators on how to interpret and *use* data to guide the selection and implementation of appropriate interventions. In this manner, professional development could be delivered where educators have the opportunity to practice using data and receive feedback on the effectiveness of their decision-making regarding intervention selection. School-based consultation models employ similar practices, wherein a consultant helps educators (e.g., consultees) identify, implement, adapt, and sustain effective practices by using data to guide the selection of interventions (Truscott et al., 2005). This would suggest that schools need training that can help educators learn how to use data effectively, in order for student change to come to fruition. Evidence for the effectiveness of teacher professional development is equivocal, with interventions that are primarily "sit and get," or lecture-based in format, exhibiting low efficacy (Joyce & Showers, 2002). In addition, previous rater training research has been limited to undergraduate participants (Chafouleas et al., 2015; Kilgus et al., 2017) or how data are collected but not used to inform intervention (von der Embse et al., 2018). Results from the present investigation show initial evidence of improved use of data to inform intervention selection; however, additional research will be necessary to inform which components are essential to facilitate the intended improvement.

Data-based decision making and accountability are one of the key domains of the National Association of School Psychologists Practice Model (NASP, 2010), with recommendations suggesting educators use a problem solving process by which to provide Tier 2 interventions to children at-risk of behavioral concerns. In order to effectively fulfill this role, schools need access to brief tools to help guide in the selection of evidence-based interventions. In addition, and in accordance with current study results, it appears educators would benefit from additional training and support to use data in an efficient and effective manner. Such support

may include coaching, with school psychologists serving as consultants directly to teachers and through participation on school-based problem-solving teams. It is incumbent upon school psychologists to critically consider the costs and benefits of training teachers to improve Tier 2 intervention selection procedures.

Limitations and Future Research

Although the findings of this study have important implications on the benefit of performance feedback on how to interpret data, there are a few study limitations that should be considered. First, the current study used vignette examples instead of having raters observe students in real time. Under typical circumstances, educators would use the ISP-Skills to rate the behavior of a student with whom they are quite familiar. However as previously discussed, it is difficult to derive true scores for all students when using rater training studies, hence the decision to use vignettes that could be compared between raters. Further, the use of vignettes and other types of analog rating settings may also be limited by the response modality (e.g., multiple choice responses), as well as potential ceiling effects regarding initial construct knowledge (see Kilgus et al., 2017). A typical school intervention selection may include many more options; thus readers are cautioned against broad generalizations. In addition, participants were not assessed for prior knowledge of or training in data interpretation. This is particularly true for use of social-emotional assessments by the individual teacher or use within the school setting. For example, some schools may frequently use social-emotional assessment tools and therefore teachers may be more familiar and comfortable using these tools (Reinke et al., 2011). Thus, the relative influence of training across participants on their data interpretation skills may be unknown.

Second, the current study only used teacher ratings of student skill levels in one context (e.g., elementary school classrooms). Future research should consider alternative raters (e.g.,

students, parents) to establish if these results are consistent across various informants in different contexts (e.g., home, social situations). Within Study 1, the sample included administrators and student support personnel; school role and context may also influence rating ability and future research should establish equivalence in rating ability. Third, although the sample was geographically diverse as it was drawn from multiple schools across two states in different geographic regions, the ISP-Skills was administered among a group of educators who were predominantly White. Future research should be conducted with a more racially diverse sample of educators and different contexts. While vignettes in Study 2 used gender neutral terminology and no demographic information, Study 1 did not which was primarily due to the longer form narrative of the vignette (as opposed to the shorter data descriptor in Study 2). Changes should include consistency in limited use of demographic information to reduce potential biases. Rating a vignette may also produce different results than rating in a naturalistic training, thus future research should examine the influence of rater training on a variety of rating scenarios (e.g., naturalistic, video, vignette).

Finally, the current study only examined the use of one teacher training with performance feedback and did not assess longer term outcomes. It is unclear to what degree additional or alternative teacher training procedures may have amplified or improved data use, or how long treatment effects may be maintained. Given the relatively small effect sizes of the current training protocols, future research will be necessary to further evaluate the retention and maintenance of the training outcomes to further inform conclusions regarding the nature and extent of appropriate training procedures.

References

Algozzine, B., Wang, C., White, R., Cooke, N., Marr, M. B., Algozzine, K., Helf, S. S., Duran, G. Z. (2012). Effects of multi-tier academic and behavior instruction on difficult-to-teach students. *Council for Exceptional Children, 79*, 45–64.

Arden, S. V., Gandhi, A. G., Edmonds, R. Z., & Danielson, L. (2017). Toward more effective tiered systems: Lessons from national implementation efforts. *Exceptional Children, 83*, 269–280. https://doi.org/10.1177/0014402917693565

Barreras, R. B. (2009). *An experimental analysis of the treatment validity of the social skills deficit model for at-risk adolescents* (Unpublished doctoral dissertation). University of California, Riverside.

Black, S. (2005). Research: good study habits are the best defense against the testing jitters. *American School Board Journal*, *192*(6), 42-44.

Bruhn, A. L., Wehby, J. H., & Hasselbring, T. S. (2019). Data-Based Decision Making for Social Behavior: Setting a Research Agenda. *Journal of Positive Behavior Interventions*, 1098300719876098.

Bundock, K., Hawken, L. S., Kladis, K., & Breen, K. (2019). Innovating the check-in, check-out intervention: A process for creating adaptations. *Intervention in School and Clinic, 55*, 169-177. https://doi.org/10.1177/1053451219842206

Castillo, J. M., Curtis, M. J., & Gelley, C. D. (2013). Gender and race in school psychology. *School Psychology Review*, *42*, 262–279.

Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2015). Preliminary investigation of the impact of a web-based module on direct behavior rating accuracy. *School Mental Health, 7*, 92-104. https://doi.org/10.1007/s12310-014-9130-z

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology, 51*, 19–57. http://dx.doi.org/10.1016/j.jsp.2012.11.001

Collaborative for Academic, Social, and Emotional Learning. (2005). *Safe and sound: An educational leader's guide to evidence-based social and emotional learning programs—Illinois edition*. Chicago: Author.

Cunningham, J. M., & Suldo, S. M. (2014). Accuracy of teachers in identifying elementary school students who report at-risk levels of anxiety and depression. *School Mental Health, 6*(4), 237-250.

Dart, E. H., & Radley, K. C. (2017). The impact of ordinate scaling on the visual analysis of single-case data. *Journal of School Psychology*, *63*, 105-118.

DiPerna, J. C. (2006). Academic enablers and student achievement: Implications for assessment and intervention services in the schools. *Psychology in the Schools, 43,* 7-17.

Eklund, K., Demarchena, S. L., Rossen, E., Izumi, J. T., Vaillancourt, K., & Kelly, S. R. (2019). Examining the role of school psychologists as providers of mental and behavioral health services. *Psychology in the Schools*, 1–13. doi:10.1002/pits.22323

Elliott, S. N., Gresham, F. M., Frank, J. L., & Beddow III, P. A. (2008). Intervention validity of social behavior rating scales: Features of assessments that link results to treatment plans. *Assessment for Effective Intervention*, *34*, 15-24.

Evans, S. W., Allen, J., Moore, S., & Strauss, V. (2005). Measuring symptoms and functioning of youth with ADHD in middle schools. *Journal of Abnormal Child Psychology*, *33*(6), 695–706.

Gresham, F. M., Elliott, S. N., & Kettler, R. J. (2011). Base rates of social skill acquisition/ performance deficits, strengths, and problem behaviors: An analysis of the Social Skills Improvement System-Rating Scales. *Psychological Assessment, 22*, 809-815.

Guion, R.M. (1965). *Personnel testing* (pp.302-353). New York: McGraw-Hill.

Hawken, L. S., & Horner, R. H. (2003). Evaluation of a targeted intervention within a schoolwide system of behavior support. *Journal of Behavioral Education*, *12*(3), 225-240.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2007) Response to intervention at school: The science and practice of assessment and intervention. In: Jimerson S.R., Burns M.K., VanDerHeyden A.M. (Eds.) *Handbook of Response to Intervention* (pp. 3-9). Boston, MA: Springer.

Joyce, B. R., & Showers, B. (2002). Student achievement through staff development (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Kilgus, S. P., Bonifay, W. E., Eklund, K., von der Embse, N. P., Peet, C., Izumi, J., Shim, H., & Meyer, L. N. (2020). *Development and validation of the Intervention Skills Profile–Skills: A brief measure of student social-emotional and academic enabling skills*. Manuscript submitted for publication.

Kilgus, S. P., Collier-Meek, M. A., Johnson, A. H., & Jaffery, R. (2014). Applied empiricism: Ensuring the validity of causal response to intervention decisions. *Contemporary School Psychology, 18*, 1-12. doi:10.1007/s40688-013-0009-z

Kilgus, S. P., Fallon, L. M., & Feinberg, A. B. (2016). Function-based modification of check-in/check-out to influence escape-maintained behavior. *Journal of Applied School Psychology, 32*(1), 24-45.

Kilgus, S. P., Kazmerski, J. S., Taylor, C. N., & von der Embse, N. P. (2017). Use of direct behavior ratings to collect functional assessment data. *School Psychology Quarterly, 32*(2), 240.

Kilgus, S. P., von der Embse, N. P., & Eklund, K. (2018). *Intervention Selection Profile –Skills.* (Unpublished measure).

Kilgus, S. P., & von der Embse, N. P. (2019). General model of service delivery for school-based interventions. In E. Dart & K. Radley (Eds.), *Handbook of Behavioral Interventions in Schools: Multi-Tiered Systems of Support* (pp. 106-133). New York, NY: Oxford University Press.

Kuchle, L. B., Edmonds, R. Z., Danielson, L. C., Peterson, A. & Riley-Tillman, T. C. (2015). The next big idea: A framework for integrated academic and behavioral intensive intervention. *Learning Disabilities Research & Practice, 30*, 150-158. doi: 10.1111/ldrp.12084

Loman, S. L., & Horner, R. H. (2014). Examining the efficacy of a basic functional behavioral assessment training package for school personnel. *Journal of Positive Behavior Interventions*, *16*, 18-30. doi:10.1177/1098300712470724

McDaniel, S. C., Bruhn, A. L., & Mitchell, B. S. (2015). A tier 2 framework for behavior identification and intervention. *Beyond Behavior*, *24*, 10–17. doi:10.1177/107429561502400103

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions, 11*, 82-93. https://doi.org/10.1177/1098300708319127

Reinke, W. M., Stormont, M., Clare, A., Latimore, T., & Herman, K. C. (2013) Differentiating

tier 2 social behavioral interventions according to function of behavior. *Journal of Applied School Psychology, 29*, 148-166. https://doi.org/10.1080/15377903.2013.778771

Rossen, E., & Von Der Embse, N. (2014). The status of school psychology graduate education in the United States. In P. L. Harrison & A. Thomas (Eds), *Best practices in school psychology: Foundations* (pp. 503–512). Bethesda, MD: National Association of School Psychologists.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-262.

Schanding Jr, G. T., & Nowell, K. P. (2013). Universal screening for emotional and behavioral problems: Fitting a population-based model. *Journal of Applied School Psychology*, *29*(1), 104-119.

Shapiro, E.S., & Heick, P.F. (2004). School psychologist assessment practices in the evaluation of students referred for social/ behavioral/emotional problems. *Psychology in the Schools, 41*, 551-561.

Sullivan, A. L., Sadeh, S., & Houri, A. K. (2019). Are school psychologists' special education eligibility decisions reliable and unbiased? A multi-study experimental investigation. *Journal of School Psychology*, *77*, 90-109. doi:10.1016/j.jsp.2019.10.006

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston, MA: Allyn & Bacon.

Truscott, S. D., Cohen, C. E., Sams, D. P., Sanborn, K. J., & Frank, A. J. (2005). The current state(s) of prereferral intervention teams: A report from two national surveys. *Remedial and Special Education*, *26*(3), 130-140.

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving decision making in school

psychology: Making a difference in the lives of students, not just a prediction about their lives. *School Psychology Review, 47*, 385–395. https://doi.org/10.17105/SPR-2018-0042.V47-4

Van Norman, E. R., & Parker, D. C. (2018). A comparison of common and novel curriculum-based measurement of reading decision rules to predict spring performance for students receiving supplemental interventions. *Assessment for Effective Intervention, 43*, 110–120. https://doi.org/10.1177/1534508417728695

Volpe, R. J., & Briesch, A. M. (2018). Establishing evidence-based behavioral screening practices in U.S. schools. *School Psychology Review, 47*, 396–402. https://doi.org/10.17105/SPR-2018-0047.V47-4

von der Embse, N. P., & Kilgus, S. P. (2018). Improving decision making: Procedural recommendations for evidenced-based assessment. *School Psychology Review, 47,* 329-332. doi:10.17105/SPR-2018-0059.V47-4

von der Embse, N. P., Kilgus, S. P., Eklund, K., Ake, E., & Levi-Neilsen, S. (2018). Training teachers to facilitate early identification of mental and behavioral health risks. *School Psychology Review*, *47*(4), 372-384.

Whitcomb, S. A. (2013). *Behavioral, social, and emotional assessment of children and adolescents*. Routledge.

Table 1

Participant Demographic Information

Demographic Variables	Study 1		Study 2	
	N	Percentage	N	Percentage
Gender				
Female	87	90.6%	181	91.4%
Male	9	9.4%	14	7.1%
Prefer not to say	-	-	3	1.5%
Race				
White	95	99.0%	159	80.3%
Black/African American	-	-	28	14.1%
American Indian or Alaskan Native	-	-	1	0.5%
Asian	-	-	1	0.5%
No Response or Prefer not to say	1	1.0%	9	4.5%
Ethnicity				
Hispanic or Latino	-	-	22	11.1%
Not Hispanic or Latino	95	99.0%	164	82.8%
No Response or Prefer not to say	1	1.0%	12	6.1%
Highest Degree Attained				
High School Diploma	-	-	1	0.5%
Associates or Technical Degree	-	-	4	2.0%
Bachelor's	43	44.8%	116	58.6%
Master's	47	49.0%	68	34.3%
Professional or Doctorate Degree	1	1.0%	9	4.5%
Certificate	4	4.2%	-	-
Other	1	1.0%	-	-
Role				
Support Staff	5	5.2%	15	7.6%
Teacher	19	19.8%	115	58.1%
ESE Teacher/Coach	-	-	36	18.2%
Student Services	-	-	13	6.6%
Administrator	1	1.0%	6	3.0%
Not Collected or Other	71	73.9%	13	6.6%
Years in Role				
Less than 1	-	-	24	12.1%
1-5 years	-	-	51	25.8%
6-10 years	-	-	43	21.7%
11-15 years	-	-	30	15.2%
16-20 years	-	-	19	9.6%
20+ years	-	-	31	15.7%
0-5 years	21	21.9%	-	-
5-10	23	24.0%	-	-
10+ years	52	54.2%	-	-

Table 2

Operational Definitions for the ISP-Skills Anchors

Skill level	Definition
Acquisition deficit	The child never displays the skill, indicating that he/she has not learned the skill.
Fluency deficit	The child only sometimes displays the skill. When he/she does display the skill, it is awkward or not in accordance with developmental expectations. The child may have learned the skill to some degree, but would benefit from additional practice to display the skill correctly.
Performance deficit	The child only sometimes displays the skill. When he/she does display the skill, it appears appropriate and in accordance with developmental expectations. However, he/she still requires additional rewards or reinforcement to display the skill.
Typical	The child displays the skill often. He/she has learned the skill and displays it at appropriate times.
Strength	The child displays the skill almost always. The skill is a strength for him/her.

Table 3

Study 1: Descriptive Statistics for Distance from "True Score" Across Groups

		Skills	
Phase	Group	*M*	*SD*
Pre-test	1	.75^	.335
	2	.84	.349
	3	.80	.414
Post-test	1	.55^	.290
	2	.55	.269
	3	.48	.265

Note. ^ variables with violations to the normality assumption. Group 1 = control, Group 2 = Basic Training, Group 3 = Basic Training + Practice.

Table 4

Study 1: Multiple Comparisons of Group Accuracy Corresponding to Repeated Measures

Comparison	Mean difference	Standard error	*p*	95% CI		Partial η^{2a}
				Lower Bound	Upper Bound	
1 vs 2	-.042	.066	1.000	-.202	.118	
1 vs 3	0.16	.068	1.000	-.150	.182	.009
2 vs 3	.058	.069	1.000	-.112	.227	

Note. Group 1 = control, Group 2 = Basic Training, Group 3 = Basic Training + Practice.

Table 5

Study 2: Descriptive Statistics for Percent Correct Across Skills Groups

Phase	Group	Skills Overall		Motivation		Academic Engagement		Study Skills		Responsible Decision Making		Relationship Skills		Self-Management		Social Awareness		Self-Awareness	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pre-test	1	.86	.11	.87	.15	.84	.16	.90	.12	.86	.24	.76	.14	.79	.22	.89	.12	.94	.11
	2	.87	.11	.86	.17	.82	.18	.90	.14	.89	.21	.77	.13	.83	.19	.92	.10	.94	.15
	3	.87	.12	.88	.15	.84	.20	.91	.13	.91	.18	.77	.13	.82	.20	.91	.11	.96	.10
Post-test	1	.87	.10	.90	.11	.92	.12	.93	.08	.82	.21	.78	.16	.86	.17	.83	.14	.92	.12
	2	.96	.07	.96	.07	.99	.05	.97	.08	.95	.09	.90	.13	.95	.14	.95	.11	.97	.06
	3	.98	.03	1.00	.02	.97	.08	.99	.03	.96	.08	.97	.07	.98	.05	.98	.06	.97	.07

Note: Values are calculated based on percent correct. Therefore, higher numbers reflect more accurate ratings. Group 1 = control, Group 2 = Basic Training, Group 3 = Basic Training + Practice.

Table 6

Study 2: Multiple Comparisons of Group Accuracy Corresponding to Repeated Measures MANOVAs at post-test

Target	Comparison	Mean difference	Standard error	*p*	95% CI		Partial η^{2a}
					Lower bound	Upper bound	
Skills - Overall	1 vs 2	-.083	.013	<.001***	-.113	-.053	.136
	1 vs 3	-.110	.012	<.001***	-.139	-.080	
	2 vs 3	-.027	.011	.064	-.054	.001	
Skills - Motivation	1 vs 2	-.07	.01	<.001***	-.10	-.04	.053
	1 vs 3	-.10	.01	<.001***	-.13	-.07	
	2 vs 3	-.03	.01	.024**	-.06	-.00	
Skills – Academic Engagement	1 vs 2	-.06	.01	<.001***	-.09	-.03	.036
	1 vs 3	-.07	.01	<.001***	-.10	-.04	
	2 vs 3	-.01	.01	1.000	-.04	.02	
Skills – Study Skills	1 vs 2	-.04	.01	<.001***	-.07	-.01	.025
	1 vs 3	-.06	.01	<.001***	-.09	-.03	
	2 vs 3	-.02	.01	.094	-.05	.00	
Skills – Responsible Decision Making	1 vs 2	-.14	.02	<.001***	-.19	-.08	.061
	1 vs 3	-.15	.02	<.001***	-.20	-.09	
	2 vs 3	-.010	.02	1.000	-.06	.04	
Skills – Relationship Skills	1 vs 2	-.12	.02	<.001***	-.173	-.07	.186
	1 vs 3	-.18	.02	<.001***	-.24	-.13	
	2 vs 3	-.07	.02	.004**	-.11	-.02	
Skills – Self-Management	1 vs 2	-.08	.02	.001***	-.14	-.03	.030
	1 vs 3	-.12	.02	<.001***	-.17	-.06	
	2 vs 3	-.04	.02	.298	-.016	.086	

Skills – Social Awareness	1 vs 2	-.12	.02	<.001***	-.16	-.07	.160
	1 vs 3	-.15	.02	<.001***	-.19	-.10	
	2 vs 3	-.03	.02	.228	-.07	.01	
Skills – Self-Awareness	1 vs 2	-.05	.02	.002**	-.090	-.015	.029
	1 vs 3	-.05	.02	.002**	-.09	-.02	
	2 vs 3	.001	.01	1.000	-.03	.04	

Note. ***p < .001, **p < .01, *p < .05. Group 1 = control, Group 2 = Basic Training, Group 3 = Basic Training + Practice.

Student C:
Ms. Summer reports that Student C often writes inappropriate words and pictures on desks and the whiteboard in the classroom and then lies about it, blaming it on peers. Student C has also been breaking school supplies and blaming it on the class gerbil.

Figure 1

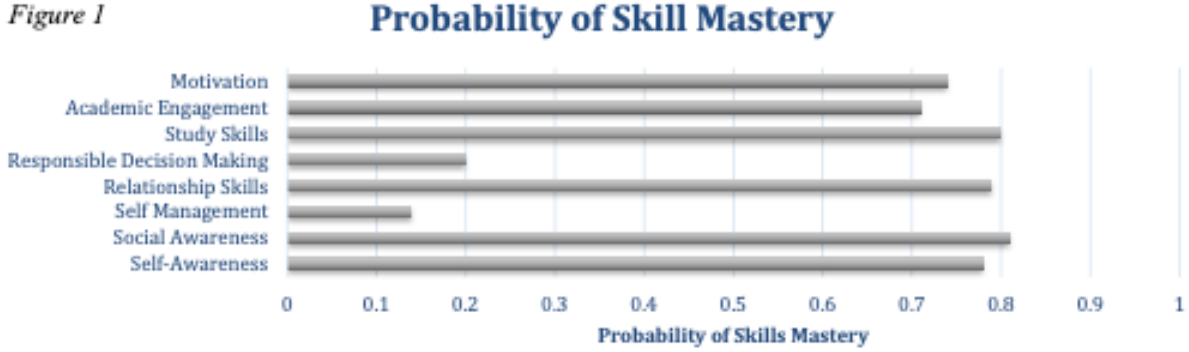

Figure 2

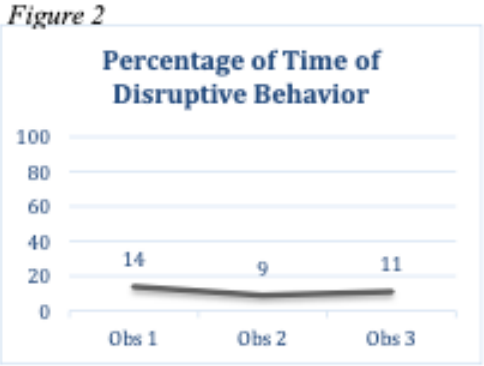

Figure 3

Check all that apply

(Note: student may have multiple or none at all in each category.)

Student has a ***deficit in***…. *(See figure 1)*

- ☐ Motivation
- ☐ Academic Engagement
- ☐ Study Skills
- ☐ Responsible Decision Making
- ☐ Relationship Skills
- ☐ Self-Management
- ☐ Social Awareness
- ☐ Self-Awareness

Student's behavior is in order to... *(See figure 2 & 3)*

- ☐ Get adult attention
- ☐ Get peer attention
- ☐ Escape/avoid something
- ☐ Get access to tangibles/activity

The appropriate next step would be to….

- ☐ Implement a skills instruction intervention
- ☐ Implement a function-based intervention
- ☐ Collect more data

Figure 1. Vignette Based Assessment of Data Use

Appendix A

Example ISP-Skills Vignette

It's Monday morning and the students in your classroom are slowly arriving for the day. At this time, students are allowed to engage in free play with their peers before structured group activities begin. Many students choose to engage in pretend play with their peers – others sit down to draw or play quietly with toys and other educational materials.

On this particular morning, you observe **Brian** walk over to two peers playing with toy farm equipment. Brian grabs a tractor from one of the boys, saying "This is my favorite," and begins to play with the toy on his own. One of the students says, "Hey, I was playing with that" and attempts to grab it back. Brian leans in to shout in the boy's face, "It's mine!!!"

On Tuesday morning, Brian enters the classroom and puts his things away. You greet Brian at his cubby and say, "Hi Brian – how are you this morning?" Brian exuberantly replies, "Great!" He then heads straight for a group of peers huddled around the class hamster's cage. Brian shoves past a peer onlooker, saying "What?! What is Hamilton the Hamster doing?"

It's Wednesday morning and Brian is one of the last to arrive. He walks in the classroom and places his things in his cubby. As he does so, you announce to the classroom, "Ok students, two more minutes until cleanup. Remember – if we are all on our best behavior today, we will go see the baby chicks hatching in Mrs. Nelson's classroom." Brian, like many other students, says "Woohoo!" He then walks over to two peers playing with puppets, saying "Hi! Can I play too?"